A lasso for hierarchical testing of interactions

Jacob Bien[*], Noah Simon [†], and Robert Tibshirani[‡]

November 7, 2012

Abstract

We consider the testing of all pairwise interactions in a two-class problem with many features. We devise a hierarchical testing framework that only considers an interaction when one or more of its constituent features has a nonzero main effect. It is based on a convex optimization framework that seamlessly considers main effects and interactions together. We provide examples— both real and simulated– that show a potential gain in power and interpretability over a standard (non-hierarchical) interaction test.

1 Introduction

We consider the standard two-class problem with $y_i = 1$ or 2 and p features $\{x_{i1}, x_{i2}, \ldots x_{ip}\}$ measured on each of $i = 1, 2, \ldots N$ observations. Large-scale hypothesis testing for the effects of individual features is a challenging problem, and has received much attention in recent years (e.g Efron (2010), Dudoit & van der Laan (2008)). The problem of testing for interactions between features is even more difficult, as there are $\binom{p}{2}$ interactions, just restricting attention to pairs of features. Buzkov, Lumley & Rice (2011) show that standard permutation tests cannot be used for interaction testing, and propose instead a parametric bootstrap-based approach. Simon & Tibshirani (2012) devise a permutation approach that exploits the close relationship between the "forward" logistic model (based on $Y|X$) and a "backward" discriminant analysis (Gaussian) model (based on $X|Y$).

When p is large, the large number of potential interactions can result in low power for detecting the true effects. One strategy used by data analysts is to first screen the data for significant main effects, and then only test for interactions among those features that are themselves significant. This can be an effective approach, but has some drawbacks. Specifically, at what threshold does one stop entering main effects? And should this threshold depend on the strength of the interactions?

The above two-stage strategy can be thought of as "hierarchical": interactions are only considered if both constituent main effects are significant. In this paper we propose a convex

[*] Departments of Biological Statistics and Computational Biology and Statistical Science, Cornell University, jbien@cornell.edu

[†] Department of Statistics, Stanford University, nsimon@stanford.edu

[‡] Departments of Health, Research & Policy, and Statistics, Stanford University, tibs@stanford.edu

formulation that models main effects and interactions together, in a hierarchical fashion. It provides a testing framework that seamlessly combines main effects and interactions. The method is closely related to the recently proposed hierarchical lasso regression method ("hierNet") of Bien, Taylor & Tibshirani (2012). We focus exclusively on pairwise interactions in the paper, but discuss possible extensions to higher order interactions in Section 7.

2 Testing interactions using a convex formulation

Before we begin, we will define some notation. Let $y_i \in \{1, 2\}$ be the class label for observation i, and denote the set of indices of the classes 1 and 2 by C_ℓ, $\ell = 1, 2$. Let x_{ij} be the jth feature for the ith observation. Furthermore let

$$\begin{aligned} E[x_{ij}] &= \mu_{j,\ell} \qquad \text{for } i \in C_\ell \\ \mathrm{Cor}\,(x_{ij}, x_{ik}) &= \frac{\sigma_{jk,\ell}}{\sigma_{j,\ell}\sigma_{k,\ell}} = \rho_{jk,\ell} \qquad \text{for } i \in C_\ell \end{aligned}$$

where $\mu_{j,\ell}, \sigma_{j,\ell}$ and $\sigma_{jk,\ell}$ are the class-specific feature means, variance and covariances and $\rho_{jk,\ell}$ is the class specific correlation.

It is reasonable to base a test of the interaction between predictors j and k on a test of equality of the correlations. In particular Simon & Tibshirani (2012) show that, in a fairly general framework, a test for equality of correlations is also implicitly a test for the interaction term in a forward logistic regression model for Y as a function of X_j and X_k.

Hence we are interested in testing the hypotheses:

$$\begin{aligned} H_{0,j} &: \mu_{j,1} = \mu_{j,2} \qquad \text{for } 1 \le j \le p \\ H_{0,jk} &: \rho_{jk,1} = \rho_{jk,2} \qquad \text{for } 1 \le j < k \le p. \end{aligned}$$

We test the main effects using the standard t-statistic

$$w_j = \frac{\bar{x}_{j,1} - \bar{x}_{j,2}}{s_j\sqrt{1/n_1 + 1/n_2}}$$

For the interactions we define a new variable:

$$z_{i,jk} = \frac{(x_{ij} - \bar{x}_{j,\ell})\,(x_{ik} - \bar{x}_{k,\ell})}{s_{j,\ell}s_{k,\ell}} \qquad \text{for } i \in C_\ell,\ \ell = 1, 2$$

where $s_{j,\ell}$ is the usual estimate of the standard deviation for covariate j in class ℓ. This new transformed variable is really just a single observation estimate of $\rho_{jk,\ell}$. As a test statistic we use the usual t statistic with this new definition of $z_{i,jk}$:

$$z_{jk} = \frac{\bar{z}_{jk,1} - \bar{z}_{jk,2}}{s_{jk}\sqrt{1/n_1 + 1/n_2}},$$

where $\bar{z}_{jk,\ell}$ is the sample mean of $z_{\cdot,jk}$ in class ℓ and s_{jk} is a pooled estimate of the standard deviation of $z_{\cdot,jk}$

We would like to select interactions based on the size of z_{jk}, but also somehow give a "boost" to interactions whose main effects are large. One could try to achieve this through a two-stage procedure, where we first screen the individual features, and then test for interactions only among those features selected at the first phase. This kind of method is explored for example in Kooperberg & LeBlanc (2008) and Hsu, Jiao, Dai, Hutter, Peters & Kooperberg (2012).

Instead, we approach this problem through a joint optimization involving both w_j and z_{jk}. Consider minimizing the function

$$\begin{aligned} L_{\lambda_1,\lambda_2}(\{\beta_j^+\},\{\beta_j^-\},\{\theta_{jk}\}) &= \frac{1}{2}\sum_{j=1}^{p}(w_j-(\beta_j^+-\beta_j^-))^2+\frac{1}{2}\sum_{j=1}^{p}\sum_{j\neq k}(z_{jk}-\theta_{jk})^2+ \\ &\quad \lambda_1\sum_{j=1}^{p}[\beta_j^+ + \beta_j^-]+\lambda_2\sum_{j=1}^{p}\sum_{k\neq j}|\theta_{jk}| \qquad (1) \end{aligned}$$

subject to the constraint that $\beta_j^{\pm} \geq 0$ (here we think of the main effect as $\hat{\beta}_j = \hat{\beta}_j^+ - \hat{\beta}_j^-$). For a fixed λ_1 and λ_2, the solutions $\hat{\beta}_j$ and $\hat{\theta}_{jk}$ are soft-thresholded versions of w_j and z_{jk}. In a sense, this criterion "selects" main effects for which $|w_j| > \lambda_1$ and interactions for which $|z_{jk}| > \lambda_2$. Such a procedure, however, does not share information between main effects and interactions. Our proposal in this paper is to add a convex hierarchy constraint to the problem, which will lead to *main-effect "informed" thresholds* for testing the interactions (and likewise interaction "informed" thresholds for testing main effects). Define

$$(\hat{\beta}^+,\hat{\beta}^-,\hat{\theta}) = \arg\min L_{\lambda_1,\lambda_2}(\{\beta_j^+\},\{\beta_j^-\},\{\theta_{jk}\}) \text{ subject to } \beta_j^{\pm}\geq 0,\ \sum_k|\theta_{jk}|\leq \beta_j^+ + \beta_j^-. \qquad (2)$$

While we could use two separate parameters, λ_1 and λ_2, we have found that $\lambda_1 = \lambda_2 = \lambda$ is a good choice, and use it throughout. Our idea is to fit a path of models (parameterized by λ) and then define the test statistic for the jkth interaction to be $\hat{\lambda}'_{jk}$, the largest λ for which either $\hat{\theta}_{jk}$ or $\hat{\theta}_{kj}$ is nonzero. In the same way, for each main effect j we compute $\hat{\lambda}_j$, the largest λ for which either $\hat{\beta}_j^+$ or $\hat{\beta}_j^-$ is non-zero. That is, letting $\hat{\beta}(\lambda) = \hat{\beta}^+ - \hat{\beta}^-$ and $\hat{\theta}(\lambda)$ denote the solution for $\lambda_1 = \lambda_2 = \lambda$, our proposed test statistics are

$$\begin{aligned} \hat{\lambda}_j &= \sup\{\lambda\geq 0:\ \hat{\beta}_j(\lambda)\neq 0\} \\ \hat{\lambda}'_{jk} &= \max\{\hat{\lambda}_{jk},\hat{\lambda}_{kj}\} \end{aligned}$$

where

$$\begin{aligned} \hat{\lambda}_{jk} &= \sup\{\lambda\geq 0:\ \hat{\theta}_{jk}(\lambda)\neq 0\} \\ \hat{\lambda}_{kj} &= \sup\{\lambda\geq 0:\ \hat{\theta}_{kj}(\lambda)\neq 0\}. \qquad (3) \end{aligned}$$

In Lemma 1 of the Appendix, we prove that (2) has a unique solution for each $\lambda > 0$, so $\hat{\lambda}_j$ and $\hat{\lambda}_{jk}$ are well-defined. Without the hierarchy constraints in (2), we would have $\hat{\lambda}_{jk} = |z_{jk}|$ and $\hat{\lambda}_j = |w_j|$. The addition of the constraint imposes a "budget" $\beta_j^+ + \beta_j^-$ on the total interactions that involve feature j. In particular, the constraint $\sum_k |\theta_{jk}| \leq \beta_j^+ + \beta_j^-$ implies that at least one of β_j^+ and β_j^- must be non-zero, in order for θ_{jk} to be non-zero. Although in theory we could have $\hat{\beta}_j^+ = \hat{\beta}_j^-$ with both values positive, this happens with probability zero under reasonable assumptions.

As a result, $\hat{\theta}_{jk} \neq 0$ implies $\hat{\beta}_j \neq 0$, and similarly for $\hat{\theta}_{kj}$. Thus the jkth interaction is nonzero if at least one of $\hat{\beta}_j$ or $\hat{\beta}_k$ is nonzero. Bien et al. (2012) (and others) call this property *weak hierarchy*, in contrast to *strong hierarchy*, which requires both $\hat{\beta}_j$ and $\hat{\beta}_k$ to be nonzero, in order for $\hat{\theta}_{jk}$ to be nonzero. In terms of our test statistics, weak hierarchy implies that

$$\hat{\lambda}'_{jk} \leq \max\{\hat{\lambda}_j, \hat{\lambda}_k\}.$$

We note also that problem (2) is convex, due to the fact that we have represented each main effect β_j as the difference of two non-negative quantities β_j^+, β_j^-. It would not be convex if we had used $|\beta_j|$ in place of $\beta_j^+ + \beta_j^-$ in the constraint above.

We call our method *convex hierarchical testing*, or sometimes "CHT" for short. The corresponding version of this proposal without the hierarchy constants, we call the *all pairs* method. This approach simply orders the z_{jk} by their absolute values.

Note also that Simon & Tibshirani (2012) find that the Fisher transform of the correlation can improve the performance of the all-pairs test. We don't use that here for CHT, since it affects the relative scaling of main effect and interaction test statistics.

Example. We simulated Gaussian data with $N = 200, p = 5$ and strong main effects for the first two variables and strong interactions for variables (1,3) and (2,5). Thus the true model obeys (weak) hierarchy. Figure 2 shows the test statistic $\hat{\lambda}$ for each main effect and interaction (horizontal axis) and the corresponding z score (w_j or z_{jk}) on the vertical axis. We see that the procedure enters variables 2, 1, interactions (1,3), variable 3 and then interaction (2,5). The latter interaction is entered before (4,5), which has a (slightly) larger z value and so would be entered first by the all-pairs procedure. This example is only for illustration, but we demonstrate later through simulations that convex hierarchical testing exhibits higher power than the all pairs approach, when the true underlying scenario is hierarchical.

3 Details of the procedure

Although the ranking of interactions from the above procedure comes from a seemingly complicated optimization problem, the solutions actually have a simple form. In particular, we show in the Appendix that

$$\hat{\theta}_{jk} \quad = S(z_{jk},\ \lambda + \hat{\alpha}_j) \tag{4}$$

Here $S(x, t) = \text{sign(x)} \cdot (|x| - t)_+$ (the soft-thresholding function), and the value $\hat{\alpha}_j \in [0, \lambda]$ emerges from the solution to problem (2), with $\hat{\alpha}_j = 0$ if the hierarchy constraint $\sum_k |\hat{\theta}_{jk}| \leq$

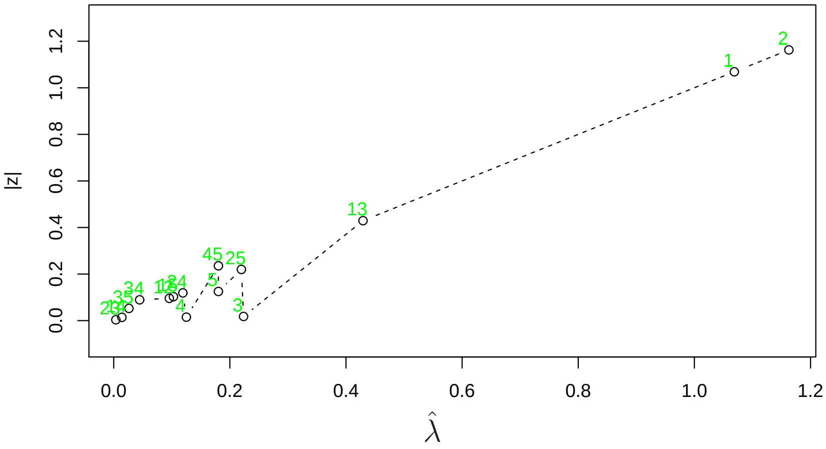


Figure 1: *Convex hierarchical testing applied to simulated example. Plot shows the test statistic $\hat{\lambda}$ for each main effect and interaction (horizontal axis) and the corresponding z score (w_j or z_{jk}) on the vertical axis.*

$\hat{\beta}_j^+ + \hat{\beta}_j^-$ is loose (i.e., a strict inequality) and greater than zero otherwise.

We interpret these conditions as follows:

- $\hat{\theta}_{jk}$ can only be non-zero if at least one of $\hat{\beta}_j^+$ and $\hat{\beta}_j^-$ is non-zero.
- if there are large main effects $\hat{\beta}_j^+$ and/or $\hat{\beta}_j^-$, then the hierarchy constraint will tend to be loose and hence $\hat{\alpha}_j = 0$. Then z_{jk} is simply soft-thresholded by λ. If this is true for all j, then CHT simply chooses all interactions z_{jk} larger than λ in absolute value. This is the same as in the all pairs approach.
- If the main effects are not very large, then the hierarchy constraint will be tight and hence $\hat{\alpha}_j$ may be larger than zero. Then from (4) we see that a higher threshold is applied to z_{jk},and it may not be selected. In this sense, a more stringent screening rule is applied to z_{jk} if its associated main effects j and k are not large.

Recall that our test statistic for interaction jk is computed as the largest value of λ such that $\hat{\theta}_{jk}$ or $\hat{\theta}_{jk}$ is nonzero. The following lemma adds justification for this test:

Property. Let $\hat{\theta}_{jk}(\lambda)$ solve (2) as a function of λ, with $\lambda_1 = \lambda_2 = \lambda$. Then $|\hat{\theta}_{jk}(\lambda)|$ is a non-increasing function of λ.

This property is proved in the Appendix (Proposition 2), and implies that once an interaction is selected, it remains selected for all smaller values of λ.

In fact, in the Appendix we are able to derive (after much work!) an explicit formula for the test statistics $\hat{\lambda}_j$ and $\hat{\lambda}_{jk}$:

Proposition. The main effect and interaction test statistics have the following closed-form expressions:

$$\begin{aligned}\hat{\lambda}_j &= \max\left\{|w_j|, \frac{|w_j| + \|z_{j\cdot}\|_\infty}{2}\right\}\\ \hat{\lambda}_{jk} &= \min\left\{|z_{jk}|, \frac{|z_{jk}|}{2} + \frac{\left[|w_j| - \sum_{l:|z_{jl}|>|z_{jk}|}\left(|z_{jl}| - |z_{jk}|\right)\right]_+}{2}\right\},\end{aligned} \tag{5}$$

where $z_{j\cdot} = \{z_{jk} : k \neq j\} \in \mathbb{R}^{p-1}$ is the vector of interaction contrasts involving the jth variable.

Proof. See Proposition 4 in the Appendix. □

These formulae are somewhat complex, but we can interpret them loosely as follows. Each main effect is "boosted" by the size of the largest interaction in its row, due to the hierarchy constraint. In contrast, each interaction is shrunken by as much as half of its size, with the shrinkage amount less when the main effect is large or the interaction is large relative to the other interactions in that row. Interestingly, λ_{jk} depends only on w_j and those interactions in the jth row that are at least as large (in absolute value) as z_{jk}. Figure 2 gives a graphical illustration of these formula. We set the number of interactions to 50. In the left panel the interaction contrasts z_j are generated as $N(0, 1)$. The plot shows the test statistic $\hat{\lambda}_j$ as a function of z_j and the main effect w (different colored curves with main effect indicated), along with the 45° line. We see that the interaction effect is shrunken substantially until it reaches about 2.75, and the amount of shrinkage is less when the main effect is larger. In the right panel there are $p - 1 = 49$ small interactions distributed as $N(0, .5^2)$ and one large interaction whose value varies along the horizontal axis. Now we see that there is shrinkage only about till a value of 1.5, and a main effect of 1.5 is sufficient to ensure no shrinkage at all.

4 A simulation study

We simulated some Gaussian data with $N = 200$ observations and $p = 50$ features in two classes $y = 1, 2$. There are three different scenarios. In the "Hierarchical truth" setup, we randomly chose five features to have strong nonzero main effects, and then 9 of their associated interactions to be nonzero, with moderate-sized effects. In the "Non-hierarchical" setup (also referred to as "No main effects") there are no main effects, and simply 45 non-zero interactions chosen at random. Finally, in the "Anti-hierarchical truth" setting, the effects are the same as in the hierarchical setup, except that the interactions are concentrated exclusively on variables having no main effects. We compared the convex hierarchical and

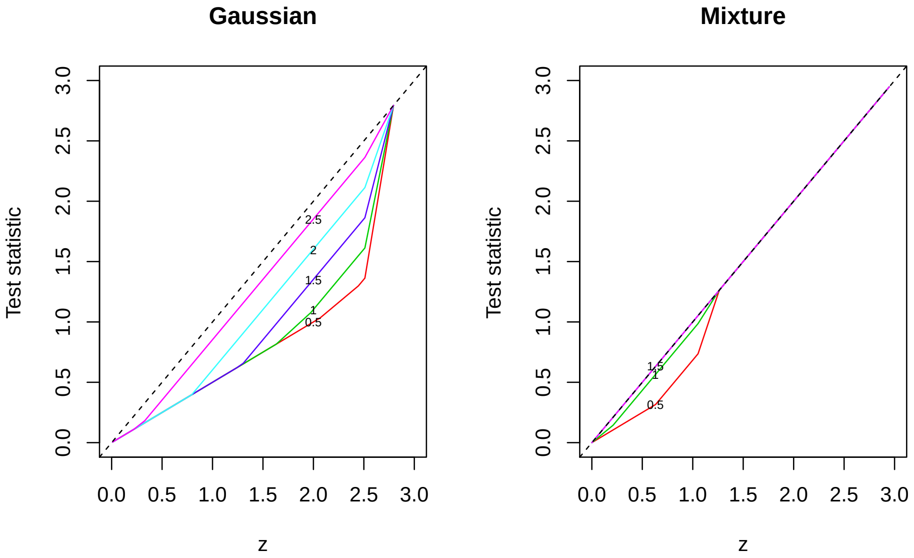


Figure 2: *Graphical illustration of formula (5) for two different distributions of interactions (two panels) and different size of main effects w (colored lines). Broken line is the $45°$ line. Full details in text.*

all-pairs testing procedures, along with two different two-stage screening methods: in the "strong" version we retained all main effects with z scores above the 75th percentile, and then in the second stage tested for interactions only among the retained variables. In the weak version, we considered all interactions among pairs of variables where ar least one variable had a z score above the 75th percentile.

Figure 3 shows the true false discovery rate for testing interactions, averaged over 20 simulations. In the hierarchical scenario, we see that convex hierarchical testing shows a substantial improvement in FDR over all-pairs, with the weak screen method performing a little worse. In the non-hierarchical scenario, CHT performs about the same as the all pairs method, while the screening methods do poorly. Not surprisingly, the false discovery rates are higher overall in the non-hierarchical scenario. In the anti-hierarchical setting, not surprisingly, convex hierarchical testing does poorly, with the weak screening perhaps a little better.

In Figure 4, we varied the strength of the main effects in the hierarchical setting and compared all-pairs and CHT in their ability to correctly detect interactions. We estimated the average number of non-null interactions called significant (over 20 replications) with FDR$<$ .2, for varying sample sizes (horizontal axis) and size of the main effect. There are 45 non-null interactions in the underlying model. In the top panel, the all-pairs method does the best, but requires close to 1000 samples to detect all 45 interactions. In the middle panel, with moderate main effects, CHT has captured all 45 effects with just 500 samples. In the hierarchical scenario, the problem is more tractable overall: with say 200 samples, CHT captures about 25 effects in the middle scenario, while the all-pairs method captures

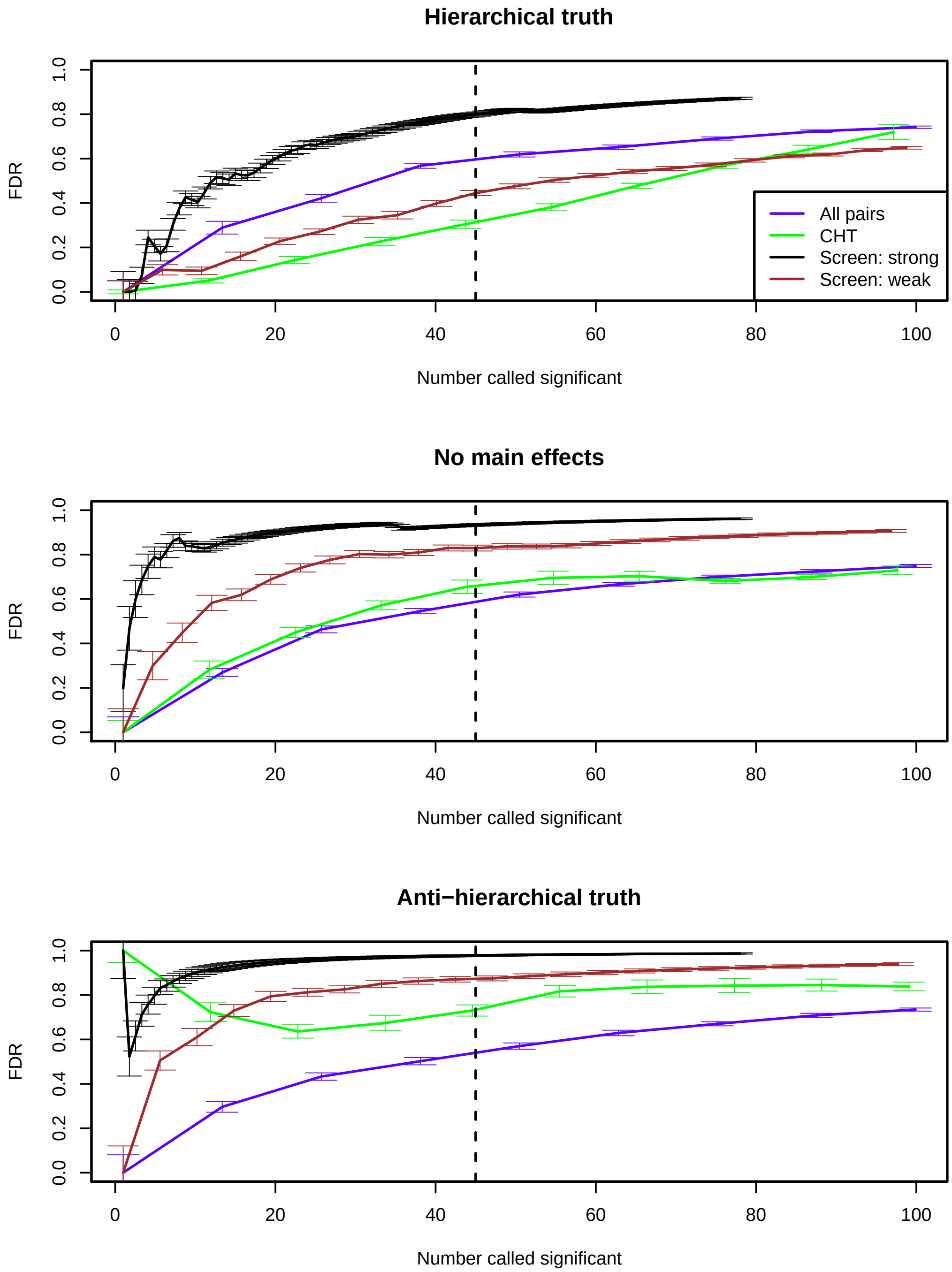


Figure 3: *True false discovery rates of four different procedures over three different settings. The convex hierarchical test has lowest FDR when the true effects are hierarchical (top panel), with the weak screen method performing a little worse. CHT performs about the same as the all pairs methods when there are no main effects (middle panel), while the screening methods do worst. In the bottom panel, there are strong main effects and interactions, but the interactions occur for predictors without main effects, and CHT performs badly. The screening methods perform poorly as well, except for weak screening for a small number of features.*

Predictor number	Name	Predictor number	Name
1	Reached menopause?	14	PTPN1i4INV
2	insulin t=-10	15	Cyp11B2x1INV
3	insulin t=60	16	PTPN1x9INV
4	insulin t=120	17	ADRB3W1R
5	HUT2SNP5	18	KLKQ3E
6	HUT2SNP7	19	AGT2R1A1166C
7	BADG16R	20	AVPR2G12E
8	AVPR2A1629G	21	MLRI2V
9	AGT2R2C1333T	22	AGTG6A
10	PPARG12	23	Cyp11B2-5paINV
11	CD36x2aINV	24	PTPN1i1
12	MLRi6INV	25	PTPN1i4
13	Cyp11B2i4INV		

Table 1: *List of predictors in the Sapphire dataset.*

only about 15 in the top panel. Increasing the size of the main effect in the bottom panel seems to make little difference.

5 Real data example- SAPPHIRe study data

This data was analyzed in Park & Hastie (2008), following the study of Huang, Lin, Narasimhan, Quertermous, Hsiung, Ho, Grove, Olivier, Ranade, Risch & Olshen (2004). The study was aimed at finding genes associated with hypertension. The dataset consists of the genotypes on 21 distinct loci and the menopausal and insulin resistant status of hypotensive and hypertensive Chinese women (with 216 and 364 subjects, respectively). The predictors are listed in Table 1.

The first four (non-genetic) predictors have the strongest effects on their own, although none were significantly different across the two groups (details not shown). Table 2 shows the first ten interactions found by the all-pairs and convex hierarchical test methods. The selected interactions are very different, with CHT focussing on clinical by genetic interaction, due to the strength of the clinical factors on their own. Figure 5 depicts the main effects and interactions found by CHT for different values of the regularization parameter λ.

To examine the stability of the two analyses, we carried out a bootstrap analysis, recording the top ten pairs appearing in the analysis from each of 500 bootstrap samples. The ten most frequently occurring interactions are shown in Table 3. We see that the CHT test shows a moderately higher reproducibility among the top ten selected interactions.

Figure 6 contains another investigation into the reproducibility of the two different methods. We divided the data many times at random into two approximately equal sized parts, and recorded how many interactions appeared in both top k lists, for each method. In the figure, k varies along the horizontal axis. Again we see that the overlap is larger for the CHT method, although this overlap is fairly small overall.

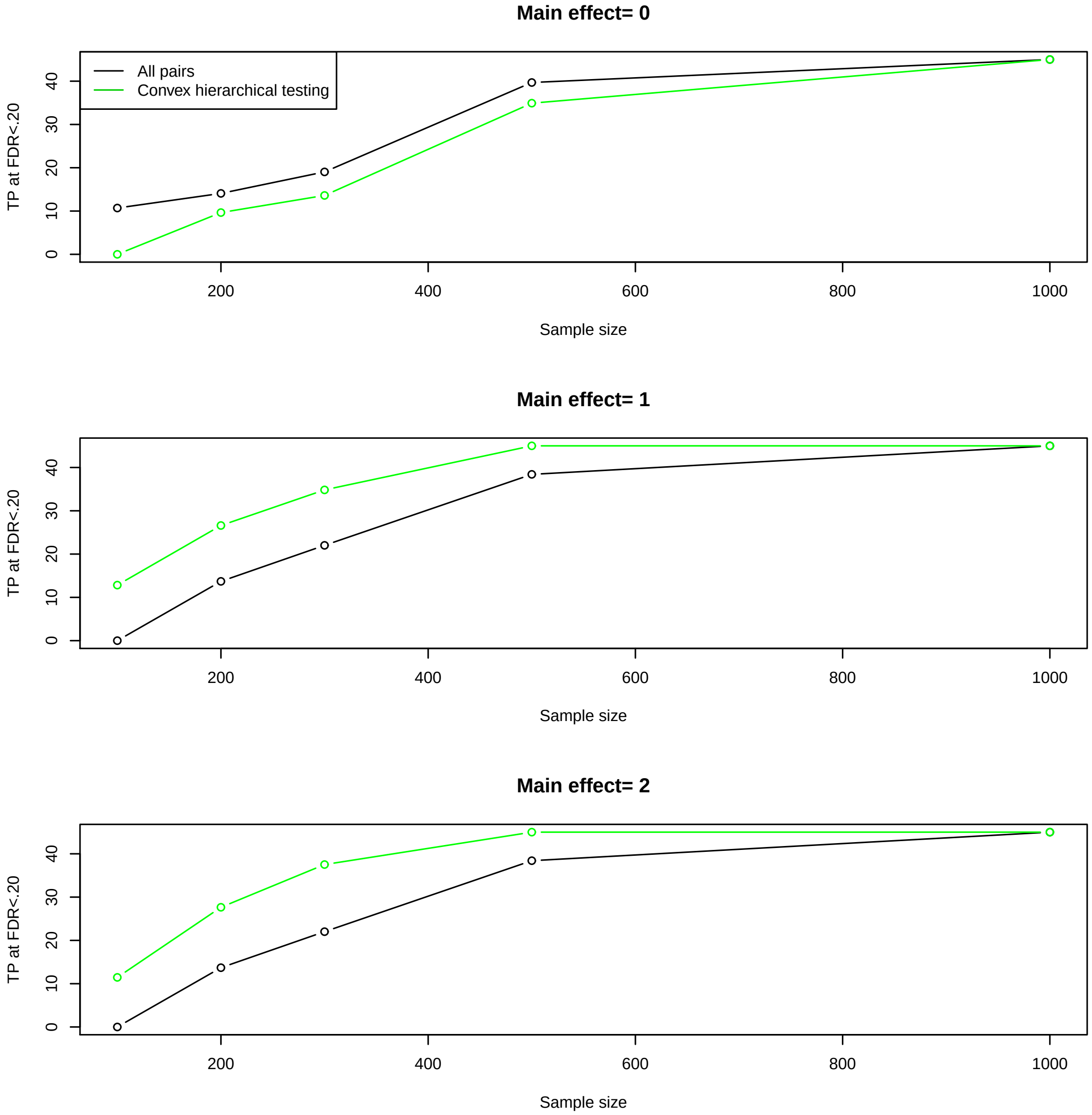


Figure 4: *Average number of True Positives (non-null interactions) called significant (over 20 replications) with FDR< .2, for varying sample sizes (horizontal axis) and size of the main effect. True number of non-null interactions is 45.*

All pairs	
Predictor 1	Predictor 2
Cyp11B2x1INV	AGTG6A
PPARG12	ADRB3W1R
PPARG12	CD36x2aINV
MLRi6INV	Cyp11B2x1INV
Cyp11B2i4INV	AGTG6A
Cyp11B2x1INV	AVPR2G12E
AGTG6A	PTPN1i1
Affected status	KLKQ3E
Reached menopause?	HUT2SNP5
AVPR2A1629G	PPARG12
Convex hierarchical test	
Reached menopause?	HUT2SNP5
insulin t=-10	MLRi6INV
Affected status	insulin t=60
Affected status	KLKQ3E
Reached menopause?	MLRi6INV
Reached menopause?	PTPN1x9INV
insulin t=-10	ADRB3W1R
Reached menopause?	AGTG6A
insulin t=60	CD36x2aINV
insulin t=60	MLRi6INV

Table 2: *Top ten interactions found by all pairs and convex hierarchical test methods.*

	All pairs	
Predictor 1	Predictor 2	Bootstrap frequency
KLKQ3E	CD36x2aINV	0.46
Cyp11B2-5paINV	PTPN1x9INV	0.43
PTPN1i4	Cyp11B2-5paINV	0.32
MLRI2V	PTPN1x9INV	0.31
MLRi6INV	CD36x2aINV	0.27
AGT2R1A1166C	Reached menopause?	0.26
PTPN1x9INV	Cyp11B2i4INV	0.25
CD36x2aINV	AGT2R2C1333T	0.21
Cyp11B2i4INV	insulin t=60	0.2
PTPN1i4INV	BADG16R	0.2
	Convex hierarchical test	
AGT2R1A1166C	Reached menopause?	0.49
HUT2SNP7	insulin t=-10	0.44
KLKQ3E	insulin t=60	0.41
insulin t=120	Reached menopause?	0.4
Cyp11B2i4INV	insulin t=-10	0.39
ADRB3W1R	insulin t=-10	0.37
Cyp11B2i4INV	insulin t=60	0.35
KLKQ3E	CD36x2aINV	0.28
Cyp11B2-5paINV	insulin t=-10	0.25
Cyp11B2i4INV	insulin t=120	0.23

Table 3: *Ten most frequent interactions found by all pairs and convex hierarchical test method over 500 bootstrap replications.*

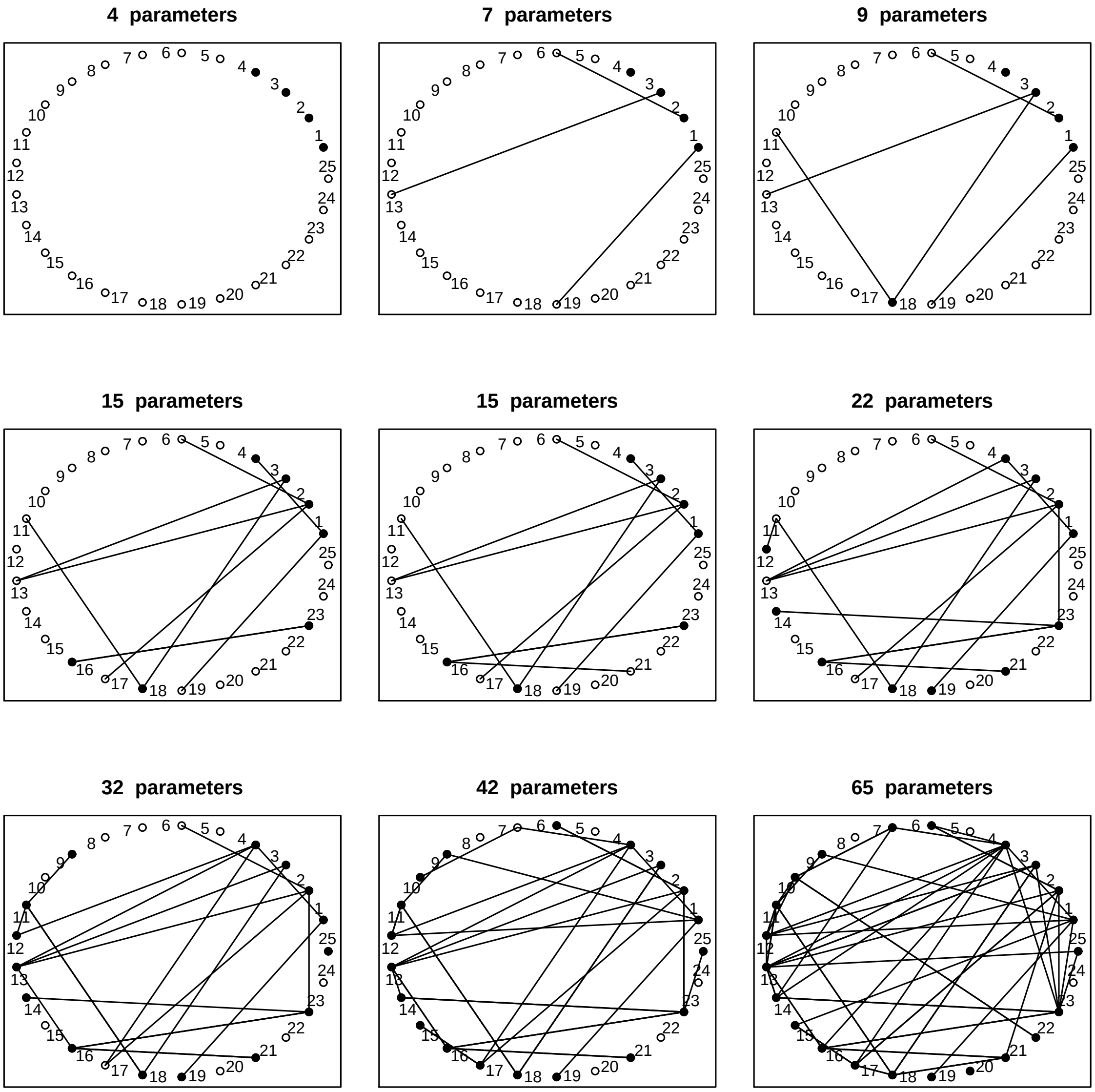


Figure 5: *Convex hierarchical testing: main effects (black dots) and interactions (edges), for 9 different decreasing values of λ. Weak hierarchy ensures that each edge is incident to at least one black dot.*

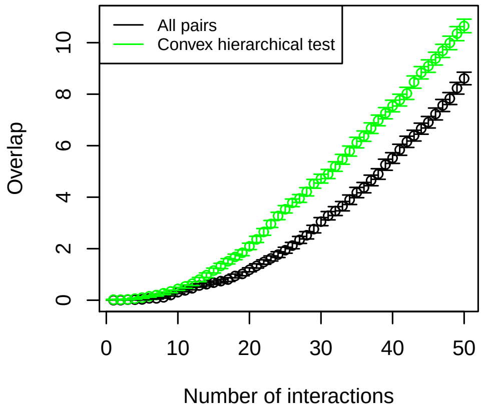


Figure 6: *Proportion of interactions found in both random halves of the data (vertical axis), as the total number of interactions selected is varied (horizontal axis).*

6 Estimation of the false discovery rate

Permutations provide a convenient and robust way to estimate false discovery rates in large-scale hypothesis testing. For example Simon & Tibshirani (2012) devise a permutation scheme for the all pairs interaction test. In this scheme, one randomly assigns a component of the interaction contrast to group 1 or group 2, by flipping the sign of the component at random.

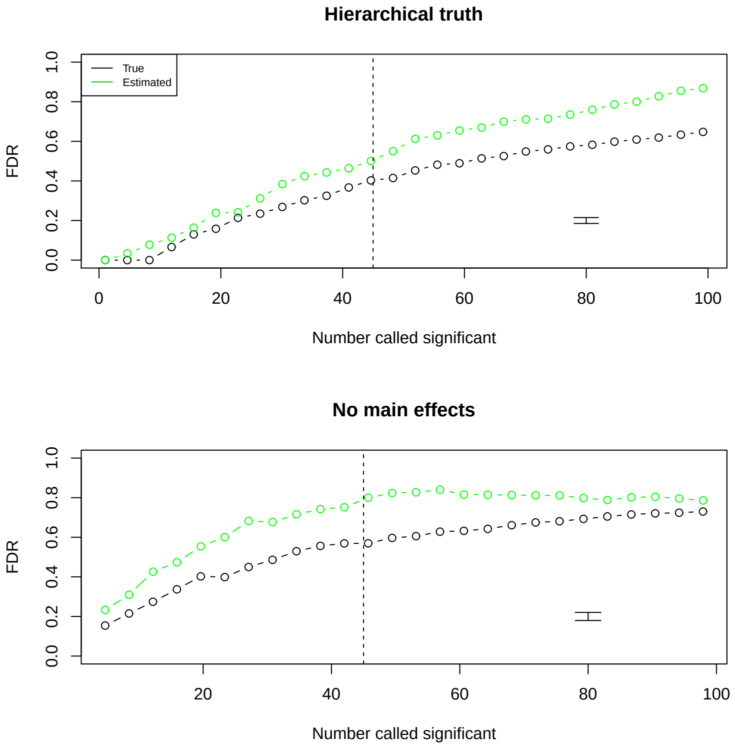
This scheme can be easily adapted to the present setting. Here are the details. The idea is to retain the main effect contrasts w_j from the original fit, and create randomized versions of the interactions as follows

$$z^*_{jk} \;=\; [\sum_{i\in C^*_2}(x_{ij}-\bar{x}_{2j})(x_{ij}-\bar{x}_{2k})/(s_{j,\ell}s_{k,\ell})-\sum_{i\in C^*_1}(x_{ij}-\bar{x}_{1j})(x_{ij}-\bar{x}_{1k})/(s_{j,\ell}s_{k,\ell})]/s_{jk}\sqrt{1/n_1+1/n_2}$$

where C^*_1, C^*_2 are the classes from a permutation y^* of the class labels y. We fit the model to the contrasts (w_j, z^*_{jk}) to obtain $\hat{\lambda}'^*_{jk,b}$, for B permutations $b = 1, 2, \ldots B$. Finally, we estimate the FDR as

$$\widehat{FDR}(\lambda) = \frac{\sum_{j,k,b} I(\hat{\lambda}'^*_{jk,b} > \lambda)/B}{\sum_{jk} I(\hat{\lambda}'_{jk} > \lambda)} \tag{6}$$

Note that this estimate pools the null distributions from all j, k pairs. This kind of pooled null is commonly used, for example in the SAM procedure (Tusher, Tibshirani & Chu 2001) and the aforementioned interaction test of Simon & Tibshirani (2012). Its accuracy is quite high in simulation studies, although we know of no rigorous results on its asymptotic properties.



Hierarchical truth

No main effects

Figure 7: *Estimation of FDR for convex hierarchical testing using permutations. Result is an average over 30 simulations, with average standard error bars shown on right. Vertical line is drawn at true number of non-zero interactions.*

Figure 7 shows the estimated FDR from this method for our example earlier. The estimate is fairly accurate, but tends to over-estimate the true FDR by a moderate amount. This may be due to the interdependence of the test statistics $\hat{\lambda}_{jk}$ for each j. In future work, it would be important to study the theoretical properties of this permutation estimate.

7 Discussion

We have proposed a hierarchical method for large-scale interaction testing, that biases its search towards interactions exhibiting at least one moderate main effect. There are a number of ways that this work could be generalized. We have focussed exclusively on pairwise interactions: extensions to k-way interactions, for $k > 2$ would bound the sum of such interactions by the size of the $k-1$ order effect. With appropriate definitions for the interaction components z_{jk}, one could also apply this procedure to interaction testing for the proportional hazards model in survival analysis.

Acknowledgements. Robert Tibshirani was supported by National Science Foundation Grant DMS-9971405 and National Institutes of Health Contract N01-HV-28183.

A A detailed look at the optimization problem (2)

A.1 Roadmap for appendix

Our procedure is based on the following optimization problem:

$$\text{Minimize} \quad \frac{1}{2}\sum_{j=1}^{p}(w_j-(\beta_j^+-\beta_j^-))^2+\frac{1}{2}\sum_{j=1}^{p}\sum_{j\neq k}(z_{jk}-\theta_{jk})^2+\lambda_1\sum_{j=1}^{p}[\beta_j^++\beta_j^-]+\lambda_2\sum_{j=1}^{p}\sum_{k\neq j}|\theta_{jk}|$$
$$\text{s.t. } \beta_j^\pm\geq 0, \sum_{k\neq j}|\theta_{jk}|\leq\beta_j^++\beta_j^- \text{ for } j=1,\dots,p.$$

Observe that this problem decouples into p separate problems, one for each j, involving the variables $(\beta_j^+,\beta_j^-,\theta_{j\cdot})$. For notational simplicity, let $w\equiv w_j$, $z\equiv(z_{j1},\dots,z_{j,j-1},z_{j,j+1},\dots,z_{jp})^T$, $\beta^\pm\equiv\beta_j^\pm$, and $\theta\equiv(\theta_{j1},\dots,\theta_{j,j-1},\theta_{j,j+1},\dots,\theta_{jp})^T$. The jth problem, whose solution is $(\hat{\beta}_j^+,\hat{\beta}_j^-,\hat{\theta}_{j\cdot})$ is

$$\text{Minimize}_{\beta^\pm\in\mathbb{R},\ \theta\in\mathbb{R}^{p-1}} \quad \frac{1}{2}(w-(\beta^+-\beta^-))^2+\frac{1}{2}\|z-\theta\|^2+\lambda_1(\beta^++\beta^-)+\lambda_2\|\theta\|_1 \quad (7)$$
$$\text{s.t. } \beta^\pm\geq 0, \|\theta\|_1\leq\beta^++\beta^-.$$

Because our original problem decouples into p separate problems involving the jth main effect and the associated $p-1$ interactions with that variable, we will study problem (7) throughout Appendix A. Once this "jth-row" problem is well-understood, we will by direct extension have studied the original problem. Therefore, for the rest of Appendix A the "inputs" to the problem will be thought of as $w\in\mathbb{R}$ and $z\in\mathbb{R}^{p-1}$ and the solution is denoted by $(\hat{\beta}^+,\hat{\beta}^-,\hat{\theta})\in\mathbb{R}^{2+(p-1)}$, which in the rest of the paper is denoted by $(\hat{\beta}_j^+,\hat{\beta}_j^-,\hat{\theta}_{j\cdot})$. In Appendix B, we apply the results of Appendix A to the paper's main problem. Here is an overview of the main elements of Appendix A:

- We write out the Karush-Kuhn Tucker (KKT) conditions for problem (7) and observe that this leads to a very simple characterization of the (primal) solution in terms of the optimal dual variables $(\hat{\gamma}^+, \hat{\gamma}^-, \hat{\alpha})$.

- We prove that the solution is unique (Lemma 1). This is important for establishing that our test statistics are well-defined (since they are defined as the largest λ for which a primal variable becomes nonzero).

- We characterize the solution path in Proposition 1. It turns out that the path is conveniently described by distinguishing among three cases: the first is the "Big main effect" case, in which $w > \|z\|_1$; the second is the "Moderate main effect" case, in which $\|z\|_\infty \leq w \leq \|z\|_1$; the third is the "Big interaction" case, in which $w < \|z\|_\infty$. The path is described in terms of some special values of λ that are defined in Lemma 2. This Lemma breaks down by case whether these values are finite and if so the relative size of these values. The description of the path in Proposition 1 is not completely closed-form, but it turns out that our characterization of the path is precise enough for our purposes.

- In particular, our test statistics only depend on when the solution paths become nonzero. Proposition 3 describes these values in closed form and Corollary 1, which is the ultimate goal of Appendix A, provides a very simple expression for the points at which the solution path becomes nonzero.

A.2 KKT conditions, a primal-dual relation, and uniqueness

The Lagrangian is

$$\begin{aligned} L(\beta^\pm, \theta; \gamma^\pm, \alpha) =& \frac{1}{2}(w - (\beta^+ - \beta^-))^2 + \frac{1}{2}\|z - \theta\|^2 + (\lambda_2 + \alpha)\|\theta\|_1 \\ &+ (\lambda_1 - \gamma^+ - \alpha)\beta^+ + (\lambda_1 - \gamma^- - \alpha)\beta^-, \end{aligned}$$

where $\gamma^\pm, \alpha \geq 0$ are dual variables corresponding to the constraints. The KKT conditions are

$$\begin{aligned} (\beta^+ - \beta^- - w) + \lambda_1 - \gamma^+ - \alpha &= 0 \\ -(\beta^+ - \beta^- - w) + \lambda_1 - \gamma^- - \alpha &= 0 \\ \theta - z + (\lambda_2 + \alpha)s &= 0 \\ \gamma^+\beta^+ &= 0 \\ \gamma^-\beta^- &= 0 \\ \alpha(\|\theta\|_1 - \beta^+ - \beta^-) &= 0 \\ \|\theta\|_1 \leq \beta^+ + \beta^-; \quad \beta^+, \beta^- &\geq 0 \\ \gamma^+, \gamma^-, \alpha &\geq 0 \end{aligned}$$

where s_k is a subgradient of $|\theta_k|$ evaluated at the solution. The stationarity conditions (first three lines) imply that

$$\hat{\beta}^+ - \hat{\beta}^- = w + (\hat{\gamma}^+ - \hat{\gamma}^-)/2$$
$$\hat{\theta} = S(z,\ \lambda_2 + \hat{\alpha})$$

Lemma 1. *Assume $\lambda_1 > 0$. Then, the solution to* (7) *is unique.*

Proof. Let $(\hat{\beta}^+, \hat{\beta}^-, \hat{\theta})$ be a solution to (7) with associated dual variables $(\hat{\alpha}, \hat{\gamma}^{\pm})$. Since (7) is convex, it suffices to show that the solution is unique in a neighborhood around this point. Noting that the objective function is strongly convex in all directions except for $(\beta^+, \beta^-, \theta)$ for which $\beta^+ - \beta^- = \hat{\beta}^+ - \hat{\beta}^-$, it remains to consider perturbations of the optimal point of the form $(\beta^+, \beta^-, \theta) = (\hat{\beta}^+ + \epsilon, \hat{\beta}^- + \epsilon, \hat{\theta} + \delta)$. If this new point is a solution, it must have a corresponding set of dual variables $(\gamma^+, \gamma^-, \alpha)$ satisfying the KKT conditions. We begin by showing in all cases that $\epsilon \neq 0$ implies $\theta = \hat{\theta}$:

- $\hat{\beta}^+ > 0, \hat{\beta}^- > 0$: In this case, $\hat{\gamma}^+ = \hat{\gamma}^- = 0$ and $\hat{\alpha} = \lambda_1$. Choosing $|\epsilon|$ small enough such that $\beta^+, \beta^- > 0$, we must have $\gamma^+ = \gamma^- = 0$ and so $\alpha = \lambda_1$. It follows that $\theta = \hat{\theta}$.

- $\hat{\beta}^+ > 0, \hat{\beta}^- = 0$: In this case, $\hat{\gamma}^+ = 0$, and $\epsilon \geq 0$ for our perturbation to be feasible. If $\epsilon > 0$, then $\beta^+, \beta^- > 0$ and so $\gamma^+ = \gamma^- = 0$ and $\alpha = \lambda_1$. Now, $\beta^+ - \beta^- = \hat{\beta}^+$, so the first KKT condition for this new point is $(\hat{\beta}^+ - w) + \lambda_1 - \alpha = 0$; however, the first KKT condition for our original point is $(\hat{\beta}^+ - w) + \lambda_1 - \hat{\alpha} = 0$. This implies that $\hat{\alpha} = \alpha$, and so $\hat{\theta} = \theta$.

- $\hat{\beta}^+ = 0, \hat{\beta}^- > 0$: Identical argument to previous case.

- $\hat{\beta}^+ = \hat{\beta}^- = 0$: Again, $\epsilon \geq 0$ is required for new point to be feasible. If $\epsilon > 0$, then $\gamma^+ = \gamma^- = 0$ and $\alpha = \lambda_1$. On the other hand, the KKT conditions for the original point implies $\hat{\alpha} \leq \lambda_1 - |w|$ (since $\pm w + \lambda_1 - \hat{\alpha} = \hat{\gamma}^{\pm} \geq 0$). Thus, $\alpha \geq \hat{\alpha}$. Now, $\hat{\theta} = 0$ means that $\|z\|_\infty \leq \lambda_2 + \hat{\alpha} \leq \lambda_2 + \alpha$, and so $\theta = 0$ as well.

Thus, our new point is $(\beta^+, \beta^-, \theta) = (\hat{\beta}^+ + \epsilon, \hat{\beta}^- + \epsilon, \hat{\theta})$, and so the objective value at the two points differs by $2\lambda_1\epsilon$. It follows that if $\epsilon \neq 0$, then one of the points is not optimal. Therefore, (7) has a unique solution. □

A.3 Characterizing the path

We take $\lambda_1 = \lambda_2 = \lambda$, and now consider the path of solutions generated by varying $\lambda > 0$. Note that by Lemma 1, it makes sense to speak of "the" path. We assume, without loss of generality, that $w \geq 0$. In studying the path, as we do in the proof of Proposition 1, we find that there are three "special" values of λ at which the behavior of the path changes. These are the following:

$$\tilde{\lambda}_1 = \min\{\lambda \geq 0 : \|S(z, \lambda)\|_1 + \lambda \leq w\},$$
$$\tilde{\lambda}_2 = \max\{\lambda \geq 0 : \|S(z, \lambda)\|_1 + \lambda \leq w\},$$
$$\tilde{\lambda}_3 = \max\{\lambda \geq 0 : \|S(z, 2\lambda)\|_1 \geq w\}.$$

In Lemma 2, we collect some facts about $\tilde{\lambda}_1, \tilde{\lambda}_2, \tilde{\lambda}_3$ that will be useful in the rest of the appendix; however, most readers may prefer to skip directly to Proposition 1, which describes the solution path.

Lemma 2. *Defining $|z_1| \geq |z_2| \geq \cdots$, the following statements hold:*

1. *$\|z\|_\infty \leq w$ iff. $\tilde{\lambda}_1, \tilde{\lambda}_2$ are finite, in which case $\tilde{\lambda}_1 \leq |z_2|$ and $\tilde{\lambda}_2 = w$.*
2. *$\|z\|_1 \geq w > 0$ iff. $\tilde{\lambda}_3$ finite, in which case $\tilde{\lambda}_3 \leq (1 - w/\|z\|_1)\|z\|_\infty/2$.*
3. *If $\|z\|_\infty \leq w \leq \|z\|_1$, then $\tilde{\lambda}_3 \leq \tilde{\lambda}_1 \leq \tilde{\lambda}_2$.*

Proof. The function $f_1(\lambda) = \|S(z, \lambda)\|_1 + \lambda$ is piecewise-linear, convex, and minimized on $\lambda \in [|z_2|, |z_1|]$ with minimal value $|z_1|$ (where $|z_1| \geq |z_2| \geq \cdots$). Thus, $[\tilde{\lambda}_1, \tilde{\lambda}_2]$ defines a nonempty interval iff. $\|z\|_\infty \leq w$. If this holds, then $\tilde{\lambda}_1 \leq |z_2|$ and $\tilde{\lambda}_2 = w$. Likewise, $\tilde{\lambda}_3$ is finite iff. $\|z\|_1 \geq w$.

The function $f_2(\lambda) = \|S(z, 2\lambda)\|_1$ is a piecewise linear, decreasing convex function with $f_2(0) = \|z\|_1$ and $f_2(\|z\|_\infty/2) = 0$. Since f_2 is convex, it lies beneath the line $L(\lambda) = \|z\|_1 - 2(\|z\|_1/\|z\|_\infty)\lambda$ on the interval $\lambda \in [0, \|z\|_\infty/2]$ and so $\tilde{\lambda}_3 \leq \tilde{\lambda}$ where $L(\tilde{\lambda}) = w$, i.e. $\tilde{\lambda} = (1 - w/\|z\|_1)\|z\|_\infty/2$.

Finally, observe that $f_1(\lambda) - f_2(\lambda) = \|S(z, \lambda)\|_1 - \|S(z, 2\lambda)\|_1 + \lambda \geq \lambda$ since $\|S(z, \cdot)\|_1$ is a non-increasing function and $2\lambda \geq \lambda$. Thus, for $\lambda > 0$, $f_1(\lambda) > f_2(\lambda)$ from which it follows that $\tilde{\lambda}_3 < \tilde{\lambda}_1$. We can only have equality if $\tilde{\lambda}_1 = \tilde{\lambda}_3 = 0$, in which case $\|z\|_1 = \|z\|_\infty = w$ (i.e. z has no more than one nonzero value), which is a 0 probability event under most reasonable models. □

Proposition 1. *Let $\tilde{\lambda}_1, \tilde{\lambda}_2, \tilde{\lambda}_3$ be as defined above, let $\tilde{\lambda}_4 = (w + \|z\|_\infty)/2$, and write $\hat{\beta} = \hat{\beta}^+ - \hat{\beta}^-$. The solution path of (7) depends on the relative size of the main effects and interactions and is given by the following:*

1. *Big main effect: $\|z\|_\infty \leq \|z\|_1 < w$.*

$$\begin{array}{rlll} \lambda \geq w \implies & \hat{\beta} = 0, & \hat{\theta} = 0 & \text{[Case III]} \\ \tilde{\lambda}_1 \leq \lambda < w \implies & \hat{\beta} = w - \lambda, & \hat{\theta} = S(z, \lambda) & \text{[Case I(i)]} \\ \lambda < \tilde{\lambda}_1 \implies & \hat{\beta} = w - \lambda + \hat{\alpha}(\lambda), & \hat{\theta} = S(z, \lambda + \hat{\alpha}(\lambda)),\ \hat{\alpha}(\lambda) > 0 & \text{[Case I(ii)(a)C.]} \end{array}$$

2. *Moderate main effect: $\|z\|_\infty \leq w \leq \|z\|_1$.*

$$\begin{array}{rlll} \lambda \geq w \implies & \hat{\beta} = 0, & \hat{\theta} = 0 & \text{[Case III]} \\ \tilde{\lambda}_1 \leq \lambda < w \implies & \hat{\beta} = w - \lambda, & \hat{\theta} = S(z, \lambda) & \text{[Case I(i)]} \\ \tilde{\lambda}_3 \leq \lambda < \tilde{\lambda}_1 \implies & \hat{\beta} = w - \lambda + \hat{\alpha}(\lambda), & \hat{\theta} = S(z, \lambda + \hat{\alpha}(\lambda)),\ \hat{\alpha}(\lambda) > 0 & \text{[Case I(ii)(a)B.]} \\ \lambda < \tilde{\lambda}_3 \implies & \hat{\beta} = w, & \hat{\theta} = S(z, 2\lambda) & \text{[Case II]} \end{array}$$

3. *Big interaction:* $w < \|z\|_\infty \le \|z\|_1$.

$$\begin{array}{rlll} \lambda \ge \tilde\lambda_4 \implies & \hat\beta = 0, & \hat\theta = 0 & \text{[Case III]} \\ \tilde\lambda_3 \le \lambda < \tilde\lambda_4 \implies & \hat\beta = w - \lambda + \hat\alpha(\lambda), & \hat\theta = S(z, \lambda + \hat\alpha(\lambda)),\ \hat\alpha(\lambda) > 0 & \text{[Case I(ii)(b) + (a)A.]} \\ \lambda < \tilde\lambda_3 \implies & \hat\beta = w, & \hat\theta = S(z, 2\lambda) & \text{[Case II]} \end{array}$$

Proof. We partition the solution path as follows:

Case I. $\hat\beta^+ > 0, \hat\beta^- = 0$.

Case II. $\hat\beta^+ > 0, \hat\beta^- > 0$.

Case III. $\hat\beta^+ = 0, \hat\beta^- = 0$.

Case IV. $\hat\beta^+ = 0, \hat\beta^- > 0$.

By Lemma 1, this is a well-defined partition of the path (i.e., for a given $\lambda > 0$, only one of the following cases holds).

Case I. $\hat\beta^+ > 0, \hat\beta^- = 0$.

The KKT conditions imply that $\hat\gamma^+ = 0$, $\hat\beta^+ = w - \lambda + \hat\alpha$, and $\hat\gamma^- = 2(\lambda - \hat\alpha)$. The conditions $\hat\gamma^-, \hat\alpha \ge 0$ and $\hat\beta^+ > 0$ require that

$$0 \le \hat\alpha \le \lambda \qquad \hat\alpha > \lambda - w$$

and, defining $f_\lambda(\alpha) = \|S(z, \lambda + \alpha)\|_1 - w + \lambda - \alpha$, they also require that

$$f_\lambda(\hat\alpha) \le 0 \text{ with } \hat\alpha f_\lambda(\hat\alpha) = 0.$$

Thus, this case occurs iff. ("if" direction by Lemma 1) there exists $\hat\alpha$ satisfying these constraints. We consider two subcases:

(i) $\hat\alpha = 0$. Requires $\lambda - w < 0$ and $\|S(z, \lambda)\|_1 + \lambda \le w$. In light of Lemma 2, this subcase happens iff. $\|z\|_\infty \le w$ and $\lambda \in [\tilde\lambda_1, w)$.

(ii) $\hat\alpha > 0$. Requires $f_\lambda(\hat\alpha) = 0$. Now, f_λ is a strictly decreasing function, so to show that such an $\hat\alpha$ exists with $[\lambda - w]_+ < \hat\alpha \le \lambda$, it suffices to check that

$$\begin{cases} \text{(a) } f_\lambda(0) > 0 \ge f_\lambda(\lambda) & \text{if } \lambda - w < 0 \\ \text{(b) } f_\lambda(\lambda - w) > 0 \ge f_\lambda(\lambda) & \text{if } \lambda - w \ge 0 \end{cases}$$

Now, $0 \ge f_\lambda(\lambda) = \|S(z, 2\lambda)\|_1 - w$, so by Lemma 2, this constraint is satisfied for all λ if $\|z\|_1 < w$ and for $\lambda \ge \tilde\lambda_3$ when $\|z\|_1 \ge w$.

(a) $f_\lambda(0) > 0 \iff \|S(z,\lambda)\|_1 + \lambda > w$, which (in light of Lemma 2) is satisfied for all λ when $\|z\|_\infty > w$ and for $\lambda \notin [\tilde\lambda_1, \tilde\lambda_2]$ when $\|z\|_\infty \le w$. Incorporating this with the constraints from $0 \ge f_\lambda(\lambda)$ and $\lambda < w$ (and using that $\|z\|_\infty \le \|z\|_1$ and $\tilde\lambda_2 = w$) gives the following cases in which (a) holds:

A. $\|z\|_\infty > w$, $\tilde\lambda_3 \le \lambda < w$,

B. $\|z\|_\infty \le w \le \|z\|_1$, $\tilde\lambda_3 \le \lambda < \tilde\lambda_1$

C. $\|z\|_1 < w$, $\lambda < \tilde\lambda_1$

(b) We have $\lambda \ge w$ and $0 < f_\lambda(\lambda - w) = \|S(z, 2\lambda - w)\|_1$, which holds as long as $\|z\|_\infty > 2\lambda - w$, i.e. $\lambda < (\|z\|_\infty + w)/2$.

A. $\|z\|_1 < w \le \lambda < (\|z\|_\infty + w)/2$. But this case can never occur since it would imply $\max\{w, \|z\|_1\} < (\|z\|_\infty + w)/2 \le (\|z\|_1 + w)/2$.

B. $\|z\|_1 \ge w$, $\lambda \ge \tilde\lambda_3$, $w \le \lambda < (\|z\|_\infty + w)/2$,

Case II. $\hat\beta^+ > 0, \hat\beta^- > 0$.

In this case, $\hat\gamma^+ = \hat\gamma^- = 0$, so we have $\hat\beta^+ - \hat\beta^- = w$ and $\hat\alpha = \lambda$. Now, $\hat\beta^+ = w + \hat\beta^- > w$ and so $\hat\beta^+ + \hat\beta^- > w$. The only remaining requirement of the KKT conditions is that $\|S(z,2\lambda)\|_1 = \hat\beta^+ + \hat\beta^- > w$. This case occurs iff. (again, we use Lemma 1 for the "if" direction) $\|S(z,2\lambda)\|_1 > w$, which by Lemma 2 is equivalent to $\|z\|_1 \ge w$ and $\lambda < \tilde\lambda_3$ (note that $\lambda = \tilde\lambda_3$ could only occur when $w = 0$, a case we exclude).

Case III. $\hat\beta^+ = 0, \hat\beta^- = 0$.

In this case, $\hat\gamma^\pm = \lambda \pm w - \hat\alpha \ge 0$, and so we require $0 \le \hat\alpha \le \lambda - w$ with $\|S(z, \lambda + \hat\alpha)\|_1 \le 0$. Putting this together gives

$$[\|z\|_\infty - \lambda]_+ \le \hat\alpha \le \lambda - w.$$

This interval is nonempty iff. $[\|z\|_\infty - \lambda]_+ \le \lambda - w$, i.e.

$$\lambda \ge \max\{w, (\|z\|_\infty + w)/2\}.$$

It is interesting to break this $\hat\beta^+ = \hat\beta^- = 0$ case into two subcases: the case in which the everything is zero even without the hierarchy constraint and the case in which the hierarchy constraint is the "reason" for everything being zero (in other words $\hat\alpha > 0$ and we wouldn't be in this case if $\hat\alpha = 0$). The hierarchy-active case occurs when $\|z\|_\infty - \lambda > 0$. In light of the lower bound on λ given above, active hierarchy sets everything to zero specifically when $\|z\|_\infty > w$ and $\lambda \ge (\|z\|_\infty + w)/2$.

Case IV. $\hat\beta^+ = 0, \hat\beta^- > 0$.

The KKT conditions imply that $\hat\gamma^- = 0$, $\hat\beta^- = \hat\alpha - (w + \lambda)$ and $\hat\gamma^+ = 2(\lambda - \hat\alpha) \ge 0$. Putting these together gives $\hat\alpha \le \lambda$ and $\hat\beta^- < -w \le 0$. In other words, this case does not occur!

We summarize these cases more succinctly:

Case I. $\hat\beta^+ > 0, \hat\beta^- = 0$.

(i) $\hat\alpha = 0$ iff. $\|z\|_\infty \le w$ and $\tilde\lambda_1 \le \lambda < w$.

(ii) $\hat\alpha > 0$ iff.

(a) A. $\|z\|_\infty > w$, $\tilde\lambda_3 \le \lambda < w$,
B. $\|z\|_\infty \le w \le \|z\|_1$, $\tilde\lambda_3 \le \lambda < \tilde\lambda_1$
C. $\|z\|_1 < w$, $\lambda < \tilde\lambda_1$

(b) $\|z\|_1 \ge w$, $\max\{w, \tilde\lambda_3\} \le \lambda < (\|z\|_\infty + w)/2$,

Case II. $\hat\beta^+ > 0, \hat\beta^- > 0$.

iff. $\|z\|_1 \ge w$ and $\lambda < \tilde\lambda_3$ (note that $\lambda = \tilde\lambda_3$ could only occur when $w = 0$, a case we exclude).

Case III. $\hat\beta^+ = 0, \hat\beta^- = 0$.

iff.

$$\lambda \ge \max\{w, (\|z\|_\infty + w)/2\}.$$

The hierarchy-active case occurs iff. $\|z\|_\infty > w$ and $\lambda \ge (\|z\|_\infty + w)/2$.

Case IV. $\hat\beta^+ = 0, \hat\beta^- > 0$.

Does not occur!

By rearranging these cases depending on the relative sizes of w, $\|z\|_\infty$, and $\|z\|_1$, we get the paths given in the Proposition statement. □

Proposition 2. *The solutions $|\hat\theta_j(\lambda)|$ are non-increasing in λ.*

Proof. Based on Proposition 1, this statement is immediate except when $\hat\alpha > 0$, i.e. Case I(ii). Referring back to this case in the proof of Proposition 1, let $\hat\alpha(\lambda)$ be the unique[1] point for which $f_\lambda(\hat\alpha(\lambda)) = 0$. Defining $t(\lambda) = \lambda + \hat\alpha(\lambda)$, we have

$$f_\lambda(\hat\alpha(\lambda)) = 0 \iff \|S(z, t(\lambda)\|_1 - t(\lambda) + \lambda = w.$$

This must hold for λ over the range in which $\hat\alpha(\lambda) > 0$ (as specified in Proposition 1). Since $\|S(z,t)\|_1 - t$ is a strictly decreasing function of t, it follows that increasing λ requires an increase in $t(\lambda)$ for λ. Since $\hat\theta = S(z, t(\lambda))$, this proves that $|\hat\theta_j|$ is nonincreasing in λ. □

[1] Since f_λ is strictly decreasing, it has a unique root.

A.4 Where do the paths become nonzero?

The results from the previous section provide sufficient information for us to derive an exact, closed-form expression for the point in the path at which each coefficient becomes nonzero.

Proposition 3. *Recall that* $\tilde{\lambda}_1 = \min\{\lambda \geq 0 : \|S(z,\lambda)\|_1 + \lambda \leq w\}$. *The main effects and interactions become nonzero at*

$$\begin{aligned} \hat{\nu} &= \sup\{\lambda : |\hat{\beta}| \neq 0\} \\ \hat{\nu}_k &= \sup\{\lambda : |\hat{\theta}_k| \neq 0\}, \end{aligned}$$

which have the following "closed form" values:

1. *Big main effect:* $\|z\|_\infty \leq \|z\|_1 < w$.

$$\begin{aligned} \hat{\nu} &= w \\ \hat{\nu}_k &= \begin{cases} |z_k| & \text{for } |z_k| \geq \tilde{\lambda}_1 \\ (|z_k| + w - \|S(z,|z_k|)\|_1)/2 & \text{for } |z_k| < \tilde{\lambda}_1 \end{cases} \end{aligned}$$

2. *Moderate main effect:* $\|z\|_\infty \leq w \leq \|z\|_1$.

$$\begin{aligned} \hat{\nu} &= w \\ \hat{\nu}_k &= \begin{cases} |z_k| & \text{for } |z_k| \geq \tilde{\lambda}_1 \\ |z_k|/2 + [w - \|S(z,|z_k|)\|_1]_+/2 & \text{for } |z_k| < \tilde{\lambda}_1 \end{cases} \end{aligned}$$

3. *Big interaction:* $w < \|z\|_\infty \leq \|z\|_1$.

$$\begin{aligned} \hat{\nu} &= (w + \|z\|_\infty)/2 \\ \hat{\nu}_k &= |z_k|/2 + [w - \|S(z,|z_k|)\|_1]_+/2 \end{aligned}$$

Proof. The expressions for $\hat{\nu}$ and $\hat{\nu}_k$ follow from Proposition 1. The only parts that are not immediate are the cases in which $\hat{\alpha}(\lambda) > 0$.

We begin by considering the "Big main effect" when $\lambda < \tilde{\lambda}_1$. By definition of $\hat{\nu}_k$ and by the expression for $\hat{\theta}$ in this case, we see that $|z_k| = \hat{\nu}_k + \hat{\alpha}(\hat{\nu}_k)$. We also know that $f_{\hat{\nu}_k}(\hat{\alpha}(\hat{\nu}_k)) = 0$ (since $\hat{\alpha}(\hat{\nu}_k) > 0$). Putting these together gives $\|S(z,|z_k|)\|_1 - w + 2\hat{\nu}_k - |z_k| = 0$, which we solve for $\hat{\nu}_k$. We also require that $\hat{\alpha}(\hat{\nu}_k) > 0$ and $\hat{\nu}_k \geq 0$, i.e. that $0 \leq \hat{\nu}_k < |z_k|$. Notice that $|z_k| < \tilde{\lambda}_1$ implies that $|z_k| > w - \|S(z,|z_k|)\|_1$, establishing that $\hat{\nu}_k < |z_k|$ for $|z_k| < \tilde{\lambda}_1$. In the "Big main effect" case, it is easy to see that $(|z_k| + w - \|S(z,|z_k|)\|_1)/2 > 0$ (which implies that $\hat{\nu}_k \geq 0$, as required). This completes the proof for the "Big main effect" case.

In the "Moderate main effect" case, if $|z_k| < \tilde{\lambda}_1$, we know that one of two cases can occur. The logic proceeds identically to the "Big main effect" case, except that we are not guaranteed that $(|z_k| + w - \|S(z,|z_k|)\|_1)/2 \geq 0$, which must hold for us to have $\hat{\nu}_k \geq 0$.

If this does hold, we're done. Assume that instead $(|z_k| + w - \|S(z, |z_k|)\|_1)/2 < 0$. This implies that $\|S(z, |z_k|)\|_1 > w$ or equivalently that $|z_k|/2 < \tilde{\lambda}_3$. This means we're in Case II and so $\hat{\nu}_k = |z_k|/2$.

Finally, we consider the "Big interaction" case. By Proposition 1, it is clear that $\hat{\nu}_k \leq \tilde{\lambda}_4$ and the expression for $\hat{\nu}_k$ will depend on whether $\hat{\nu}_k \geq \tilde{\lambda}_3$. As above, if $\hat{\nu}_k \geq \tilde{\lambda}_3$, then we would have $\hat{\nu}_k = (|z_k| + w - \|S(z, |z_k|)\|_1)/2$, which to be valid must fall within $[0, |z_k|)$ (to ensure that $\hat{\nu}_k \geq 0$ and $\hat{\alpha}(\hat{\nu}_k) > 0$). These requirements simplify to $|z_k| > w - \|S(z, |z_k|)\|_1$ and $|z_k| \geq -(w - \|S(z, |z_k|)\|_1)$. This first inequality always holds in this case since, as noted in the proof of Proposition 1, $f_1(\lambda) = \|S(z, \lambda)\|_1 + \lambda \geq \|z\|_\infty > w$. If the second inequality holds, then $\hat{\nu}_k = (|z_k| + w - \|S(z, |z_k|)\|_1)/2$. Otherwise, we have $\|S(z, |z_k|)\|_1 > w$, implying that $|z_k|/2 < \tilde{\lambda}_3$ and we're in Case II, i.e. $\hat{\nu}_k = |z_k|/2$. □

Remarks:

- In the "Big main effect" case, only small interactions are modified. In particular, the test statistic of these smallest interactions are made smaller.
- In the "Moderate main effect" case, interactions below a certain threshold are modified, with the very smallest ones being reduced to half their size.
- In the "Big interaction" case, the main effect statistic receives a positive boost and the interactions are reduced. Notice that for $|z_k| = \|z\|_\infty$, we have $\hat{\nu}_k = \hat{\nu}$, i.e. the largest interaction enters at the same time as the main effect.

Corollary 1. *The $\hat{\nu}$ and $\hat{\nu}_k$ defined as in Proposition 3 are given by*

$$\hat{\nu} = \max\left\{w, \frac{w + \|z\|_\infty}{2}\right\}$$
$$\hat{\nu}_k = \min\left\{|z_k|, \frac{|z_k|}{2} + \frac{[w - \|S(z, |z_k|)\|_1]_+}{2}\right\}$$

Proof. For the "Big main effect" and "Moderate main effect" cases, observe that $|z_k| > |z_k|/2 + [w - \|S(z, |z_k|)\|_1]_+/2$ is equivalent to $\|S(z, |z_k|)\|_1 + |z_k| > w$ (assuming $z_k \neq 0$), which can be related directly to $|z_k| < \tilde{\lambda}_1$. For the "Big interaction" case, $\|S(z, |z_k|)\|_1 + |z_k| \geq \|z\|_\infty > w$ and so $|z_k| > |z_k|/2 + [w - \|S(z, |z_k|)\|_1]_+/2$ holds automatically. □

Remarks:

- The main effect statistic is boosted when there's at least one large interaction.
- Interaction statistics get reduced by at most a factor of 2. The size of the reduction for $\hat{\nu}_k$ depends only on the size of the main effect and on those interactions that are *larger* (in absolute value) than $|z_k|$. In particular, $\|S(z, |z_k|)\|_1 = \sum_{k:|z_k|>|z_k|}(|z_k| - |z_k|)$ measures how much larger the other interactions are compared to $|z_k|$, and the larger this value, the more $\hat{\nu}_k$ is reduced.

B Test statistics

In Appendix A, we studied what happens to the jth variable (i.e. its main effect and $p-1$ interactions). In this appendix, we use the result of Appendix A to obtain a closed-form expression for our test statistics. To do so, we return to the main paper's notation, in which w_j (rather than w) represents the jth main effect, z_{jk} (rather than z_k) represents the jkth interaction, $\hat{\lambda}_j$ (rather than ν) represents the test statistic for the jth main effect and $\hat{\lambda}_{jk}$ (rather than ν_k) represents the test statistics for the jkth interaction.

Proposition 4. *Problem 2 has a unique solution path,* $\left(\hat{\beta}^+(\lambda), \hat{\beta}^-(\lambda), \hat{\theta}(\lambda)\right) \in \mathbb{R}^{2p+p(p-1/2)}$ *for* $\lambda > 0$*. The values*

$$\begin{aligned}\hat{\lambda}_j &= \sup\{\lambda \geq 0 : \ \hat{\beta}_j^+(\lambda) - \hat{\beta}_j^-(\lambda) \neq 0\} \\ \hat{\lambda}_{jk} &= \sup\{\lambda \geq 0 : \ \hat{\theta}_{jk}(\lambda) \neq 0\}\end{aligned}$$

can be expressed in closed-form as

$$\begin{aligned}\hat{\lambda}_j &= \max\left\{|w_j|, \frac{|w_j| + \|z_{j\cdot}\|_\infty}{2}\right\} \\ \hat{\lambda}_{jk} &= \min\left\{|z_{jk}|, \frac{|z_{jk}|}{2} + \frac{[|w_j| - \|S(z_{j\cdot}, |z_{jk}|)\|_1]_+}{2}\right\},\end{aligned}$$

where $z_{j\cdot} \in \mathbb{R}^{p-1}$ *is the vector of interaction contrasts involving the* j*th variable and* w_j *is the main effect contrast.*

Proof. This follows immediately from Corollary 1. □

References

Bien, J., Taylor, J. & Tibshirani, R. (2012), A lasso for hierarchical interactions.

Buzkov, P., Lumley, T. & Rice, K. (2011), 'Permutation and parametric bootstrap tests for gene-gene and gene-environment interactions', *Ann. Hum. Genet.* **1**, 36–45.

Dudoit, S. & van der Laan, M. J. (2008), *Multiple Testing Procedures with Applications to Genomics*, Springer.

Efron, B. (2010), *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, Institute of Mathematical Statistics Monographs, Cambridge University Press.

Hsu, L., Jiao, S., Dai, J. Y., Hutter, C., Peters, U. & Kooperberg, C. (2012), 'Powerful cocktail methods for detecting genome-wide gene-environment interaction', *Genetic Epidemiology* **36**(3), 183–194.
URL: *http://dx.doi.org/10.1002/gepi.21610*

Huang, J., Lin, A., Narasimhan, B., Quertermous, T., Hsiung, C. A., Ho, L.-T., Grove, J. S., Olivier, M., Ranade, K., Risch, N. J. & Olshen, R. A. (2004), 'Tree-structured supervised learning and the genetics of hypertension', *Proceedings of the National Academy of Sciences* **101**, 10529–10534.

Kooperberg, C. & LeBlanc, M. (2008), 'Increasing the power of identifying gene gene interactions in genome-wide association studies', *Genetic Epidemiology* **32**(3), 255–263.
URL: *http://dx.doi.org/10.1002/gepi.20300*

Park, M. Y. & Hastie, T. (2008), 'Penalized logistic regression for detecting gene interactions', *Biostatistics* **9**(1), 30–50.
URL: *http://biostatistics.oxfordjournals.org/content/9/1/30.abstract*

Simon, N. & Tibshirani, R. (2012), A permutation approach to testing interactions in many dimensions, Technical report, Stanford University.

Tusher, V., Tibshirani, R. & Chu, G. (2001), 'Significance analysis of microarrays applied to transcriptional responses to ionizing radiation', *Proc. Natl. Acad. Sci. USA.* **98**, 5116–5121.